MILITARY PSYCHOLOGY, 2003, *15*(1), 59–76

Expanding the Training Evaluation Criterion Space: Cross Aircraft Convergence and Lessons Learned From Evaluation of the Air Force Mission Ready Technician Program

Winston Bennett, Jr.
Human Effectiveness Directorate
Air Force Research Laboratory
Mesa, Arizona

George M. Alliger and Erik R. Eddy
The Group for Organizational Effectiveness
Albany, New York

Scott I. Tannenbaum
The Group for Organizational Effectiveness
Albany, New York, and
Department of Management
The University at Albany

This study reports the analyses of data collected from an evaluation effort for 2 Mission Ready Technician (MRT) training programs for C-141 transport and F-16 fighter aircraft crew chiefs. We obtained ratings from over 100 trainees in each program, as well as from their trainers and supervisors, both during training and in the field via survey. The goal of this research was to explore the criterion space set up for this evaluation. Whereas past evaluation research has explored task difficulty, frequency, and importance, this research explores an expanded criterion space, including task confidence, task performance, task difficulty, and task frequency. Descrip-

Requests for reprints should be sent to Winston Bennett, Jr., Warfighter Training Research Division, Human Effectiveness Directorate, Air Force Research Laboratory, 6030 South Kent Street, Mesa, AZ 85212–6061. E-mail: winston.bennett@williams.af.mil

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **JAN 2003**	2. REPORT TYPE **Journal Article**	3. DATES COVERED **00-01-2001 to 00-12-2002**

4. TITLE AND SUBTITLE
Expanding the Training Evaluation Criterion Space: Cross Aircraft Convergence and Lessons Learned from Evaluation of the Air Force Mission Ready Technician Program

5a. CONTRACT NUMBER
F33615-93-C-5011

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Winston Jr Bennett; George Alliger; Erik Eddy; Scott Tannenbaum

5d. PROJECT NUMBER
1123

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/HEA,Warfighter Training Research Division,6030 South Kent Street,Mesa,AZ,85212-6061

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL; AFRL/HEA

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/RHA, Warfighter Readiness Research Division, 6030 South Kent Street, Mesa, AZ, 85212-6061

10. SPONSOR/MONITOR’S ACRONYM(S)
AFRL; AFRL/RHA

11. SPONSOR/MONITOR’S REPORT NUMBER(S)
AFRL-RH-AZ-JA-2003-0001

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Published in Military Psychology, 2003, 15(1), 59-76

14. ABSTRACT
This study reports the analyses of data collected from an evaluation effort for 2 Mission Ready Technician (MRT) training programs for C-141 transport and F-16 fighter aircraft crew chiefs. We obtained ratings from over 100 trainees in each program, as well as from their trainers and supervisors, both during training and in the field via survey. The goal of this research was to explore the criterion space set up for this evaluation. Whereas past evaluation research has explored task difficulty, frequency, and importance, this research explores an expanded criterion space, including task confidence, task performance, task difficulty, and task frequency. Descriptive statistics, predictive regressions, and exploratory factor analyses are reported. We conclude that the data show a similar factor structure for both aircraft and that MRT frequency of task performance and confidence ratings are highly predictive of field performance. A major implication is that one way to optimize the effectiveness of training is to emphasize the development of trainee confidence at a relatively micro level, such as the task level.

15. SUBJECT TERMS
Training evaluation; Aircraft convergence; Lessons learned; Evaluation criterion; Mission ready technician; Criterion space; Task confidence; Task performance; Task difficulty; Task frequency

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Public Release**	**18**	

tive statistics, predictive regressions, and exploratory factor analyses are reported. We conclude that the data show a similar factor structure for both aircraft and that MRT frequency of task performance and confidence ratings are highly predictive of field performance. A major implication is that one way to optimize the effectiveness of training is to emphasize the development of trainee confidence at a relatively micro level, such as the task level.

TRAINING EVALUATION

Training evaluation is the programmatic process whereby the outcomes of training are tracked and analyzed. Data provided by the American Society for Training and Development (Bassi, Cheney, & VanBuren, 1997) affirmed that over 90% of surveyed private organizations evaluate training in some fashion. Bassi et al. reported that demonstrating training outcomes is one of the top 10 trends in human resources and will continue to be in the top 10 for at least 3 more years. If history is any guide, training evaluation will be a core concern for most organizations far into the future.

Given the perennial importance of training evaluation, it is not surprising that researchers are actively involved in efforts to understand how to improve our ability to evaluate effectively. One area of research has focused on the nature of training criteria. A substantial amount of research on this topic has been published. Some research has examined the relations among training criteria (e.g., Alliger & Janak, 1989; Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997). Other research has focused on drawing together the "big picture" about what criterion outcomes of hundreds of past training evaluation studies tell us (e.g., W. Bennett, 1995). Still other researchers have explored the possibility of new conceptualizations of training criteria (e.g., Kraiger, Ford, & Salas, 1993).

Perhaps surprisingly, much of the language of training evaluation criteria is still driven by the simple taxonomy proposed by Kirkpatrick in 1959 (Kirkpatrick, 1959a, 1959b, 1960a, 1960b). Although there is some discussion as to whether this taxonomy should be augmented (e.g., Alliger et al., 1997) or discarded in favor of new taxonomies (e.g., Kraiger et al., 1993), it seems indisputable that we need to continue to look closely at real-world training criteria in order both to understand how criteria are related to one another and to inform our choice of measures.

Accordingly, this article attempts to draw together empirically based lessons about training criteria as collected in a related series of training evaluation efforts. Specifically, we report analyses relevant to current research on training criteria drawn from our experiences with evaluation of two of the United States Air Force's Mission Ready Technician (MRT) training programs. As the name implies, the MRT program is designed to train and prepare a technician to be proficient in job tasks on the first day of assignment. A number of transport, fighter, and attack aircraft training programs have adopted the MRT approach (e.g., C-130, C-141, F-15, F-16, A-10, respectively). MRT training programs are typically longer than traditional technical training and involve training on actual operational aircraft.

RESEARCH OBJECTIVE

The goal in this article is not to report the effectiveness results of the MRT program. Rather, the goal is to explore the training criterion space that was set up as part of the effort to evaluate these programs. By *training criterion space,* we mean the criteria considered as a whole, in terms of fundamental structure and utility. Consequently, we took an extended correlational analysis approach, using correlations, factor analysis, and regression.

The training criterion space that we examine here is based on ratings of perceptions and performance from multiple sources. Naturally, no one study can include all conceivable measures that might be of interest in examining the effectiveness of training. Our objective, however, is to examine task ratings in numbers sufficient to allow us to examine the underlying structure, of which those ratings are the visible representation.

LITERATURE REVIEW ON TASK RATINGS

In much of the literature in the area of job analysis, human factors, and industrial or organizational psychology, a task survey is a tool for understanding which tasks in a job are most important for successful job performance, which are most frequently performed, and which are most difficult. In the MRT evaluation project, however, although frequency of task performance was assessed, a number of scales unusual for standard task surveys were also incorporated. These included, for example, task performance, number of times performed in training or on the job, and task confidence in training and on the job. One goal in including these variables was to begin an examination of an expanded criterion space for training evaluation.

General Discussion of the Concept and Use of Task Rating Descriptions

Task analysis is an important technique for anyone wishing to understand work. Research into task analysis has taken many different forms. Much of the research addresses how to write or categorize actual task statements. As one example, C. A. Bennett (1971) found that tasks could be sorted into cognitive, social, procedural, and physical categories. There are many other examples, and this stream of task analysis research has been extensively detailed in Fleishman and Quaintance

(1984). But task analysis can be used not only to understand work but also to assist in the evaluation of training. This is true because the assessment of the quality of performance at the task level, and various perceptions of tasks (such as difficulty and frequency), can be analyzed in terms of how effective the training is and how the training may be improved. In fact, the data gathered in this study did result in concrete training improvements, as mentioned later.

Research in the area of the actual ratings that are made for various tasks has been limited. That is, research is limited on whether and why tasks are to be rated in terms of frequency of performance, importance of performance, difficulty of performance, and so forth. It can be argued that this is, in part, because job analysis and task analysis, two areas where task ratings play a central role, have traditionally been largely atheoretical (aside from the inductive taxonomic work discussed by Fleishman & Quaintance, 1984).

Task ratings are obtained by asking job incumbents or supervisors to judge some task dimension directly. That is, the instructions may simply tell the incumbent to rate task "importance," "difficulty," or "frequency." Usually an anchored scale will be used (e.g., *very unimportant* to *very important;* Bernardin, 1988). Or task scales may use anchors that are somewhat derivative—as in using *crucial* as the highest anchor on a task-importance scale (Drauden, 1988), or *impossible* as the highest anchor on a difficulty scale (Beatty, Coleman, & Schneier, 1988).

It is remarkable how little we know about how people interpret and use task-rating scales. The scales that have traditionally been used are restricted in number, and there is little to guide the practitioner in terms of what scales to use for a given purpose (Fiegelson & Alliger, 1998)

For example, one problem that has received limited attention is the issue of redundancy among task-rating scales. Another problem relates to whether different sources of task ratings provide the same or different results. A brief review of research relating to these problems is presented as follows.

Task Dimension Redundancy Research

Sanchez and Levine (1989) found that task importance and criticality ratings may provide redundant information (mean *r* across four different jobs = .82), but ratings of relative time spent and task responsibility provide unique information. Sanchez and Fraser (1992) found that the type of job moderated some scale relations. For example, difficulty of learning for reference librarians was highly related to dimensions of importance and time spent and much less related for cruise line representatives. Hanges, Yost, and Cox (1991) collected ratings of task frequency, difficulty, and importance and found a high correlation between frequency and importance (mean *r* across raters = .86), suggesting redundancy. Difficulty ratings, however, appeared to provide some unique information in their study (mean *r*s with frequency and importance = .54 and .50, respectively). Bernardin (1988) found a moderate correlation between time spent and importance (r =.42, N = 370). Reilly and Israelski (1988), on the other hand, found a negligible correlation between task frequency and importance (r =.03, N = 119) but found more substantial relations between frequency and difficulty and importance and difficulty (rs = –.53 and .30, respectively). Using a very large sample of computer programmers, Alliger, Feinzig, Wong, and Douglas (1992) reported that ratings of task frequency and importance correlate substantially (about r = .60), but ratings of task difficulty to learn are virtually independent of importance and frequency.

Taken as a whole, the existing research on redundancy among task-rating dimensions has yielded mixed results. There is certainly some evidence for redundancy of importance and frequency ratings, but it is unclear to what extent this can be generalized.

Task Ratings and Rating Source

Some time ago, it was in vogue for researchers to examine "source effects"—whether and how task ratings are affected by who does the rating. For example, a typical research study might have examined whether manager ratings of how often job incumbents performed certain tasks differed from how often the incumbents themselves rated the tasks as being performed. One main reason for this research was to arrive at practical suggestions on who should complete task ratings.

Research has explored ratings made by high versus low performers (Conley & Sackett, 1987; Wexley & Silverman, 1978), individuals in various job functions (Dowell & Wexley, 1978; Schmitt & Cohen, 1989), respondents at different job levels (Cornelius, 1980; Smith & Hakel, 1979), different genders (Arvey, Davis, McGowen, & Dipboye, 1982; Arvey, Passino, & Lounsbury, 1977; Ferris, Fedor, Rowland, & Porac, 1985; Schmitt & Cohen, 1989), different races (Schmitt & Cohen, 1989), and experts versus nonexperts (Cornelius, DeNisi, & Blencoe, 1984). Results from these studies have suggested that ratings of task characteristics can vary depending on the source. In this study, we gathered ratings from supervisors and trainers as well as trainees.

Task Ratings Used in This Research

Given the background provided previously, Table 1 summarizes the criteria used in this research, categorized by several different major areas, specifically task performance, task demand, task confidence and task frequency, and timing. These criteria were gathered from a variety of sources, specifically MRT trainee, MRT graduate, instructor, and supervisor. As presented previously, there is strong empirical and theoretical justification for choosing these areas to examine.

TABLE 1
Task Measures by Venue, Source, and Operationalization

Task Criterion Measure Area	*Venue*	*Source*	*How Operationalized*
Frequency and timing of task performance	Schoolhouse	MRT trainee	Number of times you observed others performing the task
	Schoolhouse	MRT trainee	Number of times task was performed hands-on
	Schoolhouse	Instructor	Number of times tested
	Schoolhouse	MRT trainee	Recommended number of times to perform task in training (F-16 only)
	Field	MRT graduate	Number of times the task is performed per month
Task confidence	Schoolhouse	MRT trainee	Confidence in ability to perform task (1 to 5 scale) (F-16 only)
	Field	MRT graduate	Confidence that you can perform each task correctly the first time (1 to 5 scale)
Task performance	Schoolhouse	Instructor	How much of the task can the trainee perform (0 = not performed, 5 = can do complete task)
	Field	Supervisor	Percentage performing task
	Field	Supervisor	How much of the task can the MRT perform (0 = not performed, 5 = can do complete task)
Task demand	Schoolhouse	MRT trainee	Difficulty of task to learn (C-141 only)
	Schoolhouse	MRT trainee	Difficulty to perform (C-141 only)
	Field	MRT graduate	Time, in hours, to complete the task once without interruptions or delays
	Field	Supervisor	Percentage performing below standard
	Field	MRT graduate	Month on the job task first performed
	Field	Supervisor	Number of times performed before the airman could perform the task without supervision

Note. With two exceptions, the same information was gathered from both F-16 and C-141 personnel. Differences in information gathered were out of the control of the researchers. First, "difficulty to learn" information gathered from C-141 personnel was dropped in F-16 data collection and replaced by "recommended number of times to perform the task in training." Second, "difficulty to perform" information gathered from C-141 personnel was dropped in the F-16 data collection and replaced by "confidence in training." MRT = Mission Ready Technician.

THE EVALUATION EFFORT: DESCRIPTION AND METHOD

This research represented a comprehensive formative and summative evaluation of the innovative concept in technical training called the MRT training program. This program was developed by the Air Force Air Education and Training Command (AETC) to train entry-level aircraft maintenance technicians on key tasks associated with job performance in their first 6 months on the job. This certification training program is quite different from the "traditional" technical paradigm to the extent that training for certain job-relevant tasks should permit the graduates to perform some tasks without additional on-the-job training at their first job.

Task selection for MRT course content was accomplished by examining the previous course content and the content of the Major Command (MAJCOM) qualifications courses, and through consultation with technical school course personnel, MAJCOM representatives, and the career-field functional manager. In the case of the C-141 and the F-16, the tasks to be trained were already specified when the evaluation study was initiated. The tasks in the C-141 and F-16 questionnaires represented the tasks that currently are certified in all phases of the courses.

Thus, the MRT evaluation program had several objectives: to identify training needs and thereby promote continuous improvement of training courses through interpretation of field data; to develop routine benchmark measures that could be used for comparing training programs, one to the other; to drive critical business decisions (such as new course development or location of training—schoolhouse vs. field) by data; to assess course and field performance to gain a better understanding of the readiness of new technicians; and, finally, to grasp in more detail the nature and interrelations of the criteria chosen.

The evaluation effort involved the collection of both in-course (schoolhouse technical training) and field data from trainees, instructors, and supervisors in the F-16 and C-141 aircraft. Questionnaires that listed each task were distributed, and respondents provided necessary information as categorized in Table 1.

Information was collected in two studies. Study 1 was conducted with the C-141 aircraft (N = 177 trainees); Study 2, with the F-16 aircraft (N = 110 trainees). It is important to note that, due to factors outside the researchers' control, some of the data collection differed between the two aircraft. Differences are noted in the results sections where appropriate. The goal in conducting two separate studies was to provide evidence of consistent, stable results across aircraft.

The job of the MRT is to prepare and repair an aircraft, including engines. The C-141 is a large, four-engine transport, and the F-16 is a single-engine fighter. In the case of the C-141 trainees, there were 107 tasks to be rated; 95 were rated for the F-16. The median time since training (for the field ratings) was 6.5 months for the F-16 trainees and 8.5 months for the C-141 trainees. Over 50% of each sample worked the day shift, with 25% (F-16) and 35% (C-141) on the swing shift; the remainder of each group worked nights or "other."

Data was collected via questionnaire, either handed out (in the schoolhouse) or mailed (in the field). Questionnaires included instructions on completion and return.

STUDY 1 RESULTS:
C-141 AIRCRAFT

Both studies reported task means, intercorrelations, regression analyses, and exploratory factor analyses. Regression analyses were carried out via standard hierarchical linear regression; the factor analyses were carried out via the method of principal components, using varimax rotation. The rationale for these methods was

TABLE 2
Mean Task Ratings Identified by Source (C-141)

	Task Number									
Data Source	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Schoolhouse										
MRT instructor										
Task performance										
M	4.70	4.27	4.44	4.67	4.80	3.73	3.70	3.95	4.56	4.52
SD	0.61	0.84	0.75	0.62	0.41	0.86	0.81	0.72	0.64	0.71
No. of times tested										
M	5.50	1.19	1.11	6.19	7.36	1.39	1.33	1.75	3.25	3.86
SD	7.69	0.53	0.40	8.93	9.82	0.52	0.51	1.36	3.40	6.27
MRT student										
No. of times performed hands-on										
M	39.63	4.63	5.62	18.19	20.19	7.79	2.34	22.41	17.43	20.62
SD	18.97	5.58	6.30	14.34	14.38	12.06	5.50	18.15	12.23	13.61
No. of times observed										
M	28.90	5.82	5.63	14.08	15.23	6.82	2.89	15.26	13.64	19.00
SD	20.15	7.47	6.75	12.35	13.38	10.34	4.61	14.59	11.14	13.63
Difficulty to learn										
M	1.04	1.12	1.08	1.01	1.05	1.58	1.65	1.52	1.05	1.03
SD	0.25	0.36	0.27	0.11	0.27	0.70	0.70	0.75	0.27	0.16
Difficulty to perform										
M	1.05	1.17	1.11	1.01	1.06	1.55	1.63	1.51	1.08	1.09
SD	0.27	0.44	0.32	0.11	0.34	0.68	0.76	0.73	0.35	0.33
Field										
MRT supervisor										
% performing task	96.6	96.4	93.1	96.6	93.1	78.6	71.4	100.0	100.0	96.6
Task performance										
M	4.69	4.25	4.21	4.83	4.62	2.50	2.21	3.45	4.69	4.52
SD	1.00	1.46	1.52	0.93	1.29	1.88	1.91	1.33	0.71	1.12
% performing below standard	3.45	17.86	13.79	3.45	6.90	50.00	53.57	17.24	0.00	3.45
No. of times performed until no supervision										
M	1.11	1.00	0.97	1.00	0.96	0.92	0.96	1.00	1.07	1.00
SD	0.32	0.27	0.33	0.00	0.19	0.40	0.37	0.00	0.26	0.00
MRT graduate										
Month task first performed										
M	1.00	1.56	1.13	1.00	1.00	1.00	1.75	1.31	1.06	1.19
SD	1.00	0.00	1.15	0.81	0.00	0.00	0.63	1.57	0.95	0.25
No. of times task performed per month										
M	36.40	12.15	7.87	34.75	29.25	35.93	7.93	24.00	22.06	21.19
SD	23.48	12.58	6.93	25.07	17.45	65.52	12.70	21.46	13.96	16.27

(continued)

TABLE 2 *(Continued)*

	Task Number									
Data Source	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Time (in hr) required to perform task										
M	0.72	0.50	0.52	0.54	0.58	1.10	0.92	1.13	0.57	0.52
SD	1.26	1.30	1.27	1.31	1.36	1.98	2.07	1.85	1.27	1.32
Confidence										
M	4.94	4.57	4.52	4.93	4.90	3.44	3.26	4.16	4.85	4.89
SD	0.23	0.69	0.88	0.28	0.36	1.17	1.07	0.85	0.55	0.48

Note. Task names are as follows: 1 = Foreign object debris prevention; 2 = Inspect and operate H-1 heater; 3 = Inspect and operate NF-2; 4 = Statically ground aircraft; 5 = Inspect and position ground fire extinguisher; 6 = Maintenance data collection; 7 = Order and turn parts; 8 = Maintain aircraft records; 9 = Open/close troop door; 10 = Open/close crew entrance door. MRT = Mission Ready Technician.

that our primary goal was to understand the structure of the criterion space; in practice, this means studying the interrelation among criteria in a number of different ways.

Mean Task Ratings

Table 2 provides a portion of the 107 C-141 mean task ratings for the various task scales and identifies the data source, both in terms of location (schoolhouse or field) and source (instructor, student, supervisor, graduate). Similar task ratings were undertaken for the F-16 aircraft. Tables such as these were one of the most practical outcomes of this research. Tasks perceived as difficult to perform or learn in the schoolhouse could be identified and the training analyzed for possible improvements. Tasks that were not performed frequently in the field could be identified, and the emphasis on those tasks in schoolhouse training could be decreased. Similarly, schoolhouse training could be targeted for examination of those tasks for which field performance was not considered sufficient (either lower than acceptable mean performance or lower than acceptable percentage of graduates performing below standard).

Correlations

Reported correlations are not based on 107 individuals but on the mean ratings of 107 tasks. Therefore, each of the 107 data points is not subject to the same degree of error as in the usual case of correlations across individuals. Hence, the resulting correlations are likely to be more stable than is typically the case; for this reason, traditional tests of statistical significance do not apply.

Correlations among frequency of performance ratings. As pointed out previously, frequency of task performance and task importance are the two dimensions most commonly designed into task analysis questionnaires. In the schoolhouse, frequency of task performance was estimated by students in terms of the number of times the task was performed in that environment and the number of times the student observed others perform the task. In the field, frequency was estimated by the number of times the task was performed in the field per month (graduate ratings). The schoolhouse frequency ratings converged with the field per month ratings (*r*s = .38 and .40, respectively).

Correlations involving task difficulty. Students in the schoolhouse rated difficulty to learn task and difficulty to perform. These two measures correlated positively, as expected (*r* = .88). There are a number of relations that theoretically can be predicted to have a negative relation with these difficulty measures. Task difficulty and performance difficulty correlated negatively with task performance in the schoolhouse as rated by the supervisor (*r*s = –.47 and –.45, respectively). Similarly, the difficulty ratings correlated negatively with supervisor field performance ratings (*r*s = –.19 and –.20, respectively), and with graduate field confidence (*r*s = –.21 and –.18, respectively).

Correlation between confidence and performance. Confidence and performance in the field are related (*r* = .68). More will be said about the relation between confidence and performance in the discussion on the results for the F-16 and in the general discussion.

Exploratory Factor Analyses

We expected four factors to emerge from our exploratory factor analysis, mirroring the four a priori groupings of variables: task frequency and timing, task confidence, task performance, and task demand. That is, a rational grouping of variables suggested that there might be four underlying factors. Results of an exploratory factor analysis are somewhat consistent with this prediction, with several exceptions. First, information from the field (task confidence and task performance) seemed to group together (33% of total variance). Task demand seemed to split into two factors: task difficulty (14% of total variance) and task demand (9% of total variance). Task frequency and timing accounted for 19% of the total variance in the factor analysis.

Regression Analyses

An initial regression analysis explored the impact of the schoolhouse data on graduate performance. The multiple correlation of prediction of field task performance, using only schoolhouse variables, is maximized at about .28. However, if field "predictor" variables are included (but excluding any variables for which data are

TABLE 3
Regression Providing Maximal "Prediction" of Task Performance in the Field (C-141)

Variable	*Standardized Regression Coefficients*				ΔR^2	R^2
Block 1: Frequency and timing variables					.38	.38
Student: Mean number of times observed in schoolhouse	–.44	–.61	–.55	–.62		
Student: Mean number of times performed in schoolhouse	.34	.53	.47	.54		
Graduates: Mean number of times task performed in field per month	.63***	.43***	.44***	.47***		
Instructor: Mean number of times tested in schoolhouse	.04	–.04	–.05	–.02		
Block 2: Confidence variables					.22	.60
Graduates: Mean confidence in field		.53***	.51***	.47***		
Block 3: Difficulty variables					.01	.61
Student: Mean difficulty to learn in schoolhouse			–.02	–.09		
Student: Mean difficulty to perform in schoolhouse			–.08	–.04		
Block 4: Demand variables					.02	.63
Graduates: Mean time, in hours, required to perform task in field				–.08		
Graduates: Mean month task first performed in field				.11*		
Supervisor: Mean number of times performed in field until no supervision				.10*		

Note. The dependent variable is Supervisor: Mean task performance in field. All effect sizes greater than .20 are underlined.

$*p < .10$. $***p < .01$.

provided by the supervisor, the source of the criterion ratings), multiple correlation can be increased to a substantial .80 ($R^2 = .63$). The statistics for these analyses are found in Table 3.

Table 3 indicates that, in addition to the schoolhouse variables, two other major variables add to prediction in this sample: the number of times a task is reported by the graduate as being performed in the field and graduate task confidence in the field.

SUMMARY OF STUDY 1: C-141 AIRCRAFT

In summary, many of the correlations among key variables were in the expected direction for this investigation of the C-141 MRT training evaluation criteria. Schoolhouse factors predicted graduate job performance to some degree, but this prediction was significantly enhanced when information from the field was added to the regression analysis. Exploratory factor analysis showed that, although the loadings of variables onto factors was not perfect, they did tend to mirror our four a priori factors. These initial results suggest that the criteria of task frequency and timing, task confidence, task performance, and task demand is a worthwhile way of characterizing the criterion space.

STUDY 2 RESULTS: F-16 AIRCRAFT

Specific Criteria Collected for F-16 Aircraft

Consistent with the model presented in Table 1, information was gathered from the schoolhouse and field, from MRT trainees, instructors, supervisors, and MRT graduates in the areas of frequency and timing, task confidence, task performance, and task demand. The purpose of collecting information from this second aircraft was to provide evidence of consistent results.

Correlations

Again, similar to the C-141 aircraft, correlations are not based on 95 individuals but on the mean ratings of 95 tasks. Therefore, each individual of the 95 data points is not subject to the same degree of error as in the usual case of correlations across individuals.

Correlations among frequency of performance ratings. As pointed out previously, frequency of task performance and task importance are the two dimensions that are most commonly designed into task analysis questionnaires. The schoolhouse frequency ratings converged with the field per month ratings (*rs* = .56 and .75, respectively).

Correlations between confidence and performance. Some of the most interesting correlations involve MRT confidence. Confidence in the schoolhouse is related to performance in that environment (*r* = .85). Similarly, confidence and performance in the field are related (*r* = .86). Confidence in the schoolhouse is related to confidence in the field (*r* = .40). The fact that confidence relates to performance within the same situation (schoolhouse, field) supports Bandura's (1977) contention that efficacy-type variables are situation specific in their predictive abilities. The fact that confidence is so highly related to performance has practical implications for training.

Exploratory Factor Analyses

We expected four factors to emerge from our exploratory factor analysis, mirroring the four groupings of variables: task frequency and timing, task confidence, task performance, and task demand. Results of an exploratory factor analysis are somewhat consistent with this prediction, with several exceptions. Task frequency and timing and task demand were two clear factors in the factor analysis, accounting for 15% and 11%, respectively, of the total variance. Task confidence in the schoolhouse loaded with instructor ratings of performance (12% of variance), and confidence in the field loaded with other field performance measures (38% of the variance). This indicates a blending of our criteria with venues.

Regression Analyses

An initial regression analysis explored the impact of the schoolhouse data on graduate performance. The multiple correlation of prediction of field task performance, using only schoolhouse variables, is maximized at about .51. However, if field "predictor" variables are included (but excluding any variables for which data are provided by the supervisor, the source of the criterion ratings), multiple correlation can be increased to a substantial .88 ($R^2 = .77$). The statistics for this regression are found in Table 4.

SUMMARY OF STUDY 2:
F-16 AIRCRAFT

Again, for this aircraft, many of the correlations among key variables were in the expected direction. Schoolhouse factors predicted graduate job performance to some degree, but this prediction was significantly enhanced when information from the field was added to the regression analysis. Exploratory factor analysis showed that, although the loadings of variables onto factors were not perfect, they did tend to mirror our four factors. These initial results again support the idea that criteria of task frequency and timing, task confidence, task performance, and task demand is a useful way of thinking about the criterion space.

DISCUSSION

One major goal of the MRT project was eminently practical: to provide informative feedback about how MRT trainees are doing, both in the schoolhouse and in the field, to instructional designers, trainers, and those charged with supervision of training and design. Other briefings and documents have addressed these issues.

TABLE 4
Regression Providing Maximal "Prediction" of Task Performance in the Field (F-16)

Variable	*Standardized Regression Coefficients*			ΔR^2	R^2
Block 1: Frequency and timing variables				.65	.65
Instructor: Mean number of times tested in schoolhouse	.08	.01	.03		
Student: Mean number of times performed in schoolhouse	.27**	.01	.00		
Student: Mean number of times observed in schoolhouse	–.06	–.12	–.24		
Student: Recommended number of times performed in training	–.34	.03	.09		
Graduates: Mean number of times task performed in field per month	.66***	.14	.22		
Block 2: Confidence variables				.34	.87
Student: Confidence in training		–.13	–.11		
Graduates: Mean confidence in field		.86***	.79***		
Block 3: Demand variables				.02	.88
Graduates: Mean time, in hours, required to perform task in field			–.03		
Graduates: Mean month task first performed in field			–.13		
Supervisor: Mean number of times performed in field until no supervision			–.02		

Note. The dependent variable is Supervisor: Mean task performance in field. All effect sizes greater than .20 are underlined.
$^{**}p < .05$. $^{***}p < .01$.

However, the interested reader has only to examine Table 2 to see how mean task ratings on a variety of scales, derived from a variety of sources and from different venues, can be a valuable diagnostic tool for a training initiative. What tasks are not performed well? Which are the hardest to learn? On which tasks do trainees have the least confidence that they can perform well? These kinds of questions, answered by clear schoolhouse and field data, can inform revision of training or the design of new training. Thus, the practical goal of this MRT training evaluation effort, although formally outside the scope of this article, can be appreciated from a review of our results.

From a technical research point of view, the goals of this study were severalfold. For many years, researchers have been trying to identify underlying taxonomies of human work performance. Increasingly, too, efforts have been made to clarify the training criterion space. As the literature review showed, these are conceptually related efforts. Some training criteria, particularly on-the-job measures, are simply ways to assess work performance.

One feature of the MRT effort is to advance our understanding of the training criteria space: how training criteria relate and how measures converge and diverge. This convergence–divergence, or underlying structure, is an important issue because it addresses which measures offer unique variance to the criterion space and which are more or less overlapping. For the researcher who wants to be able to tell practitioners what to measure, this type of independence analysis can provide practical guidance.

The Structure of the Criterion Space

Exploratory factor analyses suggest a large degree of replication between the two aircraft. In each case, the largest single factor was what might be termed *field performance,* with supervisor rating of MRT graduate performance receiving the highest loadings. Also loading highly on this factor were percentage performing task and task field confidence. The second factor was composed of schoolhouse measures of frequency and timing of testing and observation. The third factor was either schoolhouse task difficulty (C-141) or task confidence (F-16). Finally, there was a field task demand (or field task difficulty).

Given that there was some difference in measures between the two aircraft, the amount of similarity in structure is very substantial. We can reasonably suggest that field performance, training frequency and timing (or duration), difficulty to learn (or demand), field demand (or difficulty), and confidence represent five replicable dimensions of task measures. Thus, when designing a training evaluation program, elements of each of these could be incorporated into the measures used.

Prediction

Prediction of field performance by schoolhouse variables. In the case of each aircraft, the strongest schoolhouse predictor of field performance was the number of times a task was reported to have been performed in the schoolhouse. This may indicate that practice and familiarity of, and duration (or frequency) of exposure to, a task can have an impact on how well that task is performed later. Indeed, although we accept as a truism that "practice makes perfect," this truism is certainly supported here.

Optimal prediction of field performance. Tables 3 and 4 report hierarchical regressions, using field performance as the criterion. As noted previously, the main structural difference between the two tables is that task difficulty variables were available for C-141. Squared multiple correlation in both cases exceeds .60—this is very high, given that common method variance can be ruled out as an explanation. If we accept only those variables as good predictors that have standardized regression weights above an arbitrary .20 in both samples, then we find the predictive variables are number of times observed in schoolhouse, number of times performed in the field, and confidence in the field. Although the first two may relate to practice (but, surprisingly, number of times observed in the schoolhouse receives a negative weight), it is clear that confidence ratings alone add over 20% of unique variance accounted for over schoolhouse variables. This relation between confidence and performance deserves some discussion.

The predictive power of confidence. The same general relation between performance and confidence holds in both aircraft. This relation is in both cases a remarkably strong one (*r*s = .68 and .86, respectively). Why should confidence predict performance so well? One obvious answer is found in the work of Bandura (1977, 1984). He proposed a concept called self-efficacy. Self-efficacy is a focused belief in the ability to perform a specific task or in a specific arena of activity. Hence, the MRT measures of confidence, because they were precisely at the task level, can be considered measures of self-efficacy in Bandura's sense. Given the strength of prediction found in this study, it is interesting to note that Bandura (1984) suggested that self-efficacy will "usurp the lion's share of the variance in human conduct" (p. 252).

Implications for Training Evaluation

There are several implications for training evaluation that can be derived from the research presented in this article. First, assessing training at the task level can result in concrete improvements in training. Second, task difficulty and timing–frequency measures can be a useful addition to the evaluator's collection of training evaluation measures. Third, because source effects (e.g., supervisor vs. trainer vs. trainee) do not obscure rationale content relations among measures, evaluators can with some confidence obtain measures from different sources without worrying that source effects will substantially impact the outcomes. A fourth practical implication may be that in training, and later in the field, fostering self-confidence at the task level is critical.

ACKNOWLEDGMENTS

This effort was supported under Contract F33615–93–C–5011 with the U.S. Air Force Research Laboratory, Human Effectiveness Directorate, Brooks Air Force Base, Texas. We thank Mark S. Teachout and Douglas J. Sego for their comments and recommendations during the conduct of this effort.

The views expressed in this article are those of the authors and do not necessarily reflect the official policies or position of their respective organizations or of the

Air Force Education and Training Command. Those interested in understanding in complete detail the applied, practical outcomes of these evaluation efforts should contact the first author.

REFERENCES

Alliger, G. M., Feinzig, S. L., Wong, W., & Douglas, P. J. (1992). *Are ratings ofjob tasks redundant? An investigation of the job of computer programmer and general considerations.* Unpublished manuscript.

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology, 42,* 331–342.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis on the relations among training criteria. *Personnel Psychology, 50,* 341–358.

Arvey, R. D., Davis, G. A., McGowen, S. L., & Dipboye, R. L. (1982). Potential sources of bias in job analytic procedures. *Academy of Management Journal, 25,* 618–629.

Arvey, R. D., Passino, E. M., & Lounsbury, J. W. (1977). Job analysis results as influenced by sex of incumbent and sex of analyst. *Journal of Applied Psychology, 62,* 411–416.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84,* 191–215.

Bandura, A. (1984). Recycling misconceptions of perceived self-efficacy. *Cognitive Therapy and Research, 8,* 231–255.

Bassi, L., Cheney, S., & VanBuren, M. (1997). Training industry trends. *Training and Development, 51,* 46–64.

Beatty, R. W., Coleman, S. C., & Schneier, C. E. (1988). Human resource planning and staffing. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol. 1, pp. 138–156). New York: Wiley.

Bennett, C. A. (1971). Toward empirical, practical, comprehensive task taxonomy. *Human Factors, 13,* 229–235.

Bennett, W., Jr. (1995). Factors that influence the effectiveness of training in organizations: A meta-analytic review. *Dissertation Abstracts International.* (UMI No. 9615769)

Bernardin, H. J. (1988). Police officer. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol. 1, pp. 1242–1254). New York: Wiley.

Conley, P. R., & Sackett, P. R. (1987). Effects of using high- versus low-performing job incumbents as sources of job-analysis information. *Journal of Applied Psychology, 72,* 434–437.

Cornelius, E. T., III. (1980). A comparison of holistic and decomposed judgment strategies in job analyses by job incumbents. *Journal of Applied Psychology, 72,* 434–437.

Cornelius, E. T., DeNisi, A. S., & Blencoe, A. G. (1984). Expert and naive raters using the PAQ: Does it matter? *Personnel Psychology, 37,* 453–464.

Dowell, B. E., & Wexley, K. N. (1978). Development of a work behavior taxonomy for first-line supervisors. *Journal of Applied Psychology, 63,* 563–572.

Drauden, G. M. (1988). Task inventory analysis in industry and the public sector. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol. 2, pp. 1051–1071). New York: Wiley.

Ferris, G. R., Fedor, D. B., Rowland, R. M., & Porac, J. K. F. (1985). Social influence and sex effects on task performance and task perceptions. *Organizational Behavior and Human Performance, 36,* 66–78.

Fiegelson, M., & Alliger, G. M. (1998, April). *A contemporary look at task analysis: The relationship among traditional and alternative task rating scales.* Paper presented at the annual convention of the Society of Industrial and Organizational Psychology, Dallas, TX.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks.* Orlando, FL: Academic.

Hanges, P. J., Yost, P. R., & Cox, J. F. (1991). *Task inventories: Do different formulas identify different tasks as critical?* Unpublished manuscript.

Kirkpatrick, D. L. (1959a). Techniques for evaluating training programs. *Journal of ASTD, 13*(11), 3–9.

Kirkpatrick, D. L. (1959b). Techniques for evaluating training programs: Part 2—Learning. *Journal of ASTD, 13*(12), 21–26.

Kirkpatrick, D. L. (1960a). Techniques for evaluating training programs: Part 3—Behavior. *Journal of ASTD, 14*(1), 13–18.

Kirkpatrick, D. L. (1960b). Techniques for evaluating training programs: Part 4—Results. *Journal of ASTD, 14*(2), 28–32.

Kraiger, K., Ford, K. J., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78,* 311–328.

Reilly, R. R., & Israelski, E. W. (1988). Telecommunications craftsworkers. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol. 2, pp. 1265–1286). New York: Wiley.

Sanchez, J. I., & Fraser, S. L. (1992). On the choice of scales for task analysis. *Journal of Applied Psychology, 77,* 545–553.

Sanchez, J. I., & Levine, E. L. (1989). Determining important tasks within jobs: A policy capturing approach. *Journal of Applied Psychology, 74,* 336–342.

Schmitt, N., & Cohen, S. A. (1989). Internal analyses of task ratings by job incumbents. *Journal of Applied Psychology, 74,* 96–104.

Smith, J. E., & Hakel, M. D. (1979). Convergence among data sources, response bias and reliability and validity of a structured job analysis questionnaire. *Personnel Psychology, 32,* 677–692.

Wexley, K. N., & Silverman, S. B. (1978). An examination of differences between managerial effectiveness and response patterns on a structured job analysis questionnaire. *Journal of Applied Psychology, 63,* 646–649.